DOCUMENT RESUME

ED 449 813 IR 058 057

AUTHOR Stach, Ron
TITLE Searching for Utopia: Best Practices in Digitizing the Corporate Library.
PUB DATE 1999-11-00
NOTE 68p.; Master's Research Paper, Kent State University.
PUB TYPE Dissertations/Theses (040)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Corporate Libraries; Digital Computers; *Information Management; Information Storage; Information Technology; Organizational Development; Preservation; Special Libraries
IDENTIFIERS Digital Collections; *Digitizing

ABSTRACT

Using an in-depth review of the literature and the author's own personal experience, this paper attempts to distill the eight best practices of those undertaking digitization projects. Results are intended specifically to assist special librarians undertaking a digitization effort in a corporate setting, but conclusions could equally apply to the public or non-profit sector. Digitization experiences are examined from public, academic, and corporate libraries, books and how-to manuals for digital librarians, and unpublished presentations from workshops and conferences. Discussion includes knowledge management and its relationship to a corporate digitization project. Questions focus on why, what and how to digitize. Major issues that should be considered in a digitization project are revealed, such as immature technologies, lack of standards, and intellectual freedom. Key conclusions are that there is no absolute "right" way to digitize, but that best practices are clearly discernible. Opportunities for further research are suggested. The best practices of effective digitizers (also listed in an appendix) are identified as follows: (1) It is best to be on the cutting edge, not the bleeding edge, of technology; (2) Knowledge management is not left out; (3) People are key, not the technology; (4) End users are considered; (5) Communication and training are a line item in the budget; (6) Effectiveness is measured, reported, and addressed; (7) Choices are made carefully and thoughtfully; and (8) Clues are left for future generations. Appendices also provide guides to scanner and Optical Character Recognition technologies, including their own select bibliographies, as well as sample project planning aids (process flow diagrams and project checklist). (Contains 58 references.) (AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 449 813

SEARCHING FOR UTOPIA: BEST PRACTICES IN DIGITIZING THE CORPORATE LIBRARY

A Master's Research Paper Proposal submitted to the
Kent State University School of Library
And Information Science
In partial fulfillment of the requirements
for the degree of Master of Library Science

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D.P. Wallace

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

by

Ron Stach

November 1999

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IR058057

BEST COPY AVAILABLE

Master's Research Paper

Ronald L. Stach

B.A., St. John's University, 1980

M.L.S., Kent State University, 1999

Approved by

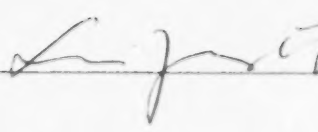
Adviser  Date 11/4/99

TABLE OF CONTENTS

I. INTRODUCTION 1
Purpose 2
Limitations 4
Definitions 5

II. REVIEW OF THE LITERATURE 7

III. METHODOLOGY 19

IV. FINDINGS AND END PRODUCTS 21

V. CONCLUSIONS 29

APPENDIX 1: Best practices 31

APPENDIX 2: Digitization project checklist 32

APPENDIX 3: Sample scanning process 37

APPENDIX 4: Elements of an image processing system 38

APPENDIX 5: Scanner guide 39

APPENDIX 6: OCR guide 49

REFERENCE LIST 57

SEARCHING FOR UTOPIA:
BEST PRACTICES IN DIGITIZING
A CORPORATE LIBRARY

CHAPTER I

INTRODUCTION

Special libraries are arguably an integral part of many corporations' success or failure. They can range in size from one person to many. They can be centralized or decentralized, depending on organizational needs. Even the name they are given – information centers, corporate libraries, or research center are but a few examples – varies, reflecting the norms of the institutional culture they represent.

The evolution of corporate librarians and the various roles they have played is reflected in the myriad titles conferred on them over the years. Librarian, cybrarian, information specialist, research specialist, knowledge manager, and more recently, knowledge worker (Markum 1998) can all be found in the literature. In corporate settings in particular, technology has been an important part of the information specialists' world. Corporate librarians usually possess developed on-line searching skills, and have often sought innovative ways to efficiently store and access information. They plan an essential role in the information and knowledge lifecycle. Librarians are trained and experienced at identifying, organizing, storing, delivering, and to an increasingly greater extent, even synthesizing information relevant to an institution's needs.

Enter the digital library. Though the term has been around since the late 1970s, only recently has it begun to once again shape the role information specialist. Web-enabled technology has driven the growth of corporate intranets, which in turn have driven the demand for digitized content. Many corporate libraries are facing the prospect that, in order to stay competitive, the systems they use to store and access internal and external knowledge also have to keep up.

Though the myth of the paperless office has largely been disproven, the topic of corporate digital libraries is "hot" judging from the recent literature and conference agendas. Instead of budgeting for additional storage space, organizations are now budgeting for additional servers with the intention of digitizing everything but the four walls and librarians themselves. But is it always cost effective to do so? Have access issues been resolved? At what point does it make sense for an organization to digitize rather than accrue and store paper? And once the decision to digitize is made, what are the considerations? In short, is the corporate digital library madness, myth, or magic?

Purpose

Simply put, this investigation primarily attempts to identify the best practices for digitizing a corporate collection. An important secondary objective is to distill some of the assumptions that might otherwise remain unchallenged when a digitizing effort is undertaken. Finally, this effort is to provide a small

bank of resources, tools, and frameworks that assist the digitizer or person overseeing the digitization effort.

Admittedly, a driver of this research was in part a personal one: A little more than a year ago, I was asked to undertake a relatively small-scale effort to digitize a paper collection for a highly knowledge-dependent organization. As I delved deeper into the project, I found an array of widely disparate sources (some with contradictory opinions) that addressed elements of such an undertaking, but very little that was comprehensive, providing the bigger picture. Those sources that attempted to develop this often ended up with a list of vague generalities, leaving important details unaddressed.

Is a list of "best practices" feasible, practical, or even desirable? Would such a list be usable by any type of corporation or for-profit organization that is planning to digitize some or all of its resources? Is there a "tool kit" of resources that can be universally used? Clearly there are arguments for and against. On the "pro" side, all organizations facing a digitization project must address key issues such as technology requirements, budgetary considerations, and training issues. On the "anti-" side, information and knowledge requirements vary widely by industry, making developing a list of standards difficult.

Still, this paper argues that such a list, however incomplete or generic, is a useful starting place for the corporate librarian faced with such an endeavor. It should be seen as the lighthouse to the port, not as the harbor itself.

Limitations

This study is neither comprehensive, nor highly quantitative. Lessons learned are derived from examining digitization projects from all types of libraries (non-profit organizations, academic institutions, public libraries, and private corporations) as well as from personal experience. However, these best practices were developed specifically to aid the corporate library and some of its unique challenges. Thus, a fundamental limitation of this investigation is that results are probably not applicable outside a business environment.

The conclusions of this investigation are not based on carefully designed primary research, and findings are arguably more subjective than in traditional fact-based analysis. Results serve, however, to identify where both the "doughnuts" and the "holes" are in a relatively new domain: the digital corporate library.

Digitizing can cover a host of resources, including text, images, graphics, numeric data, audio and video. Of these, this paper primarily addresses text and graphics for several reasons. First, technology for digitizing text resources (with embedded graphics) is the most common information resource within most businesses. Second, digitizing images (e.g., photographs, art, handwritten materials) is more common to organizations concerned with archiving and preservation or for those serving scholarly or academic purposes (e.g., museums). Finally, most corporations requiring research usually use some subscription-based information service, such as Dow Jones or Dialog. Thus, text

storage and retrieval is a model that most corporations understand in theory, if not in practice.

Although this study touches upon knowledge management and its role in the corporation, it does not dive deeply into its theory, application, or value. It is not intended to definitively demonstrate any positive correlation between successful digitization efforts and successful knowledge management. Likewise, the results should not be interpreted to mean that digitizing will increase an organization's ability to use knowledge effectively. For that purpose, the body of literature surrounding best practices in knowledge management is large and well established.

This study does not assume any specific or common research and information infrastructure within a corporate setting. The lessons learned should serve equally well the one-person librarian, a business information center, or a content-specific research services unit. The only assumption made is that there is a recognized need for enterprise-wide access to information proprietary to the organization. Companies choose to support that need in various ways. If digitizing text is one of them, this should provide guidance. If it is not one of them, this should also help determine if that decision, unconscious or conscious, needs to be reconsidered.

Definitions

The term "digital library" has not been consistently defined, though its earliest appearance seems to be about 1977 (Saffady 1995). Some sources insist

that a precise definition should be rigorously enforced, while others use the term interchangeably with "virtual" library, electronic library, or even to mean an (OPAC) online public access catalog (Feldman 1999). Saffady provides perhaps the most comprehensive picture how the term has been used in the literature, and recent usage has continued to be ambiguous.

This paper uses the definition and purpose of the digital library as developed by the Association of Research Libraries in 1995. It denotes the digital library (and terms used synonymously) as having five elements in common:

1. The digital library is not a single entity;
2. The digital library requires technology to link the resources of many
3. The linkages between the many digital libraries and information services are transparent to the end users;
4. Universal access to digital libraries and information services is a goal;
5. Digital library collections are not limited to document surrogates: they extend to digital artifacts that cannot be represented or distributed in printed formats.

"Digitizing" as used in this study refers to the process of turning non-electronic materials into an electronic format that can be retrieved and displayed. It closely resemble the term imaging, which refers to "the management of paper documents, records, forms, photos, and drawings by capturing, storing, indexing, retrieving, and distributing them electronically." (Imaging for Process Improvement: Report of the Imaging Committee 1995)

CHAPTER II

REVIEW OF THE LITERATURE

A scan of the literature on digitization efforts revealed four general categories, outlined below. Each of these were reviewed with attention to the way they addressed the following questions: (1) Why digitize? (2) What should be digitized? (3) How should it be digitized? Collectively these sources address any or all of these key questions.

Before proceeding, it should be noted that much valuable literature exists on digitizing collections other than for corporations and has been included here. While reasons for digitizing often differ substantially between the private and public sectors, the core processes of selection, conversion, and delivery are virtually identical. For that reason, the body of literature on digitizing public and academic collections cannot be ignored. Many of these institutions have already walked through the digitization mine field, and have lived to tell us what to expect.

Books, Journal Articles, Professional Opinions

A substantial body of scholarly work exists on digitization, including case studies and investigative research. Some of this literature stems from well-known and large-scale digital library efforts, such as the Library of Congress "Making of America" project. Other works document some of the technical aspects, and a few include annotated bibliographies.

Before undertaking a digitization project, institutions need to ask themselves why. For academic or public libraries, possible answers are to enhance access to local collections, increase access to remote sources, assist users in navigating an increasingly complex information environment, and to provide more information at less expense (Crawford 1999).

Philosophically, few corporate librarians would disagree. But for corporations, key drivers typically revolve around cost and benefit: Will the project result in an improved product line at reduced cost? Will it give the organization a competitive edge? Will it improve the organization's knowledge management capabilities? Sometimes the question comes down to what the cost of *not* digitizing will be.

Others assert that a corporate digital library will not be successfully if it is unbundled from a knowledge management strategy (Bonaventura 1997). This point is echoed often. Most companies have elaborate systems to capture, codify and disseminate knowledge, but the culture may not support or reward innovative knowledge sharing (Manville and Foote 1997). In addition, companies that continually "push" information in front of its employees are missing the point: the power of knowledge comes from the "pull" or need-to-know side.

A careful Knowledge Management strategy and tools that support it is only part of the picture, however. The human support infrastructure (librarians, publishers, design teams) are what one author argues will make or break a

Knowledge Management program (Bonaventura 1997). Most corporations today do not observe the same standards for knowledge currency and content that is observed in the publishing industry, but really should. Librarians are especially capable of seeing that these standards are enforced.

Knowledge Management tools can be effective, but are still relatively young in development. These tools tend to emphasize the organization and transfer of knowledge, rather than identifying and capturing it. A list of the most common technology adoptions in knowledge intensive firms may be useful when undertaking digitizing projects (Liebowitz 1999). According the Knowledge Management Consortium, the most recent findings show the mostly widely adopted technologies used by organizations that are highly knowledge dependent (e.g., consultancies, law firms):

Knowledge Management Technology Adoption	
E-mail	100%
Internet	100%
Videoconferencing	100%
Project management systems	91%
Groupware	91%
Intranet	82%
Customer management systems	73%
Skills inventory systems	64%
Yellow Pages for knowledge	44%

Why digitize? One perspective maintains that, fundamentally, corporations preserve information for the same reasons libraries do: for its intrinsic value. In this sense, "information is information, regardless of its source." (Megill 1997) Corporations need to preserve this "corporate memory" through effective (digital) information storage and retrieval. The original format of the information is immaterial; what matters is that it gets captured and digitized for further re-use. Interestingly, Megill recommends that corporations do not spend a great deal of time in retro conversions, arguing that most information in a corporation has a life span of about a month, with re-use dropping dramatically after 3 months and again after 3 years. The implication here is that choosing what to digitize should be carefully considered as the reasons why.

In one journalist's opinion, libraries see three clear benefits to digitizing: preservation, convenience (access), and space conservation. However, he also points out the technological tradeoffs that must be considered. Scanning and OCR are still not perfect processes, and the prospect of manual data entry to reproduce is not one many organizations can afford Other issues, such as copyright and lack of formatting standards, are identified. Still, the author suggests, the hazards do not outweigh the benefits (Lesk 1997).

Some identify training and the "pain of technology" as issues for the digital library. If there are a variety of interfaces (web or otherwise), the librarian must be aware of them and how to use them. While corporations sometimes do

better at training than public institutions, the need for ongoing training remains an important consideration. In addition to the hands-on application training, librarians need to know the latest technology trends and be able to identify those that might be adopted to fill specific needs. (Crawford 1999).

In corporations, the "cutting edge" technology may not be the most cost effective or offer the best solution to data and information problems (Blue 1984). An important consideration should be life span of hardware and software. One information technologist admonishes to not even try to buy hardware that will last much longer than five years, but suggests that you *do* think long term when making software decisions (Tennant 1998). Yet another suggests that librarians should not get so far ahead of their users (technologically) that they become irrelevant (Crawford 1998).

How should digitized content be made accessible? In the context of document delivery, intranets are becoming an emerging standard. Managing intranets, both their development and content, presents organizational challenges. In one documented case study of a large petroleum and exploration and production company, the thoughtful hardware and software choices enabled both searching, retrieval and editing of documents in their native formats (Fishenden 1997). Based on this, it is clear that the intended use of the documents (i.e., the basis for further revision or editing) must be a key factor in deciding the final digitized format.

Overall guiding principles or guidelines appear less frequently and tend to appear in accounts of academic digital library endeavors. In one case, it was observed that the digitizers were usually not concerned with long-term data preservation (Beagrie 1998). The report also noted the importance of using standards such as controlled vocabularies for classification of digitized materials. The best practices and standards applied to digitized objects were contingent upon four criteria, listed in descending order of importance: the data type of the resource being considered (file formats appropriate for electronic texts are different than those f or digital images), upon their ability to support a data resource in its intended use (the file format appropriate for Web-accessible image thumbnails may be different than that appropriate for digital surrogates created for the purposes of preservation), upon their cost of implementation and future maintenance, and upon available technology.

Another report prepared by an academic provided a rich summary of the processes, applications and issues involved in digitization projects (Lynn-George 1996). Some of the more salient principles identified (attributed to Donald J. Waters, Director of Library and Administrative Services at Yale University) are: (1) Think in terms of life cycles, not permanency; (2) Simplify by working on large quantities of materials with few problems before working on small quantities of materials with large problems; and (3) Adopt an incremental approach...with clear but relatively modest goals, measurable benchmarks. The implications of these for corporate libraries are clear.

Academic libraries offer other resources for digitization efforts. Columbia University, for example, provides an online source of technical recommendations for digital imaging projects. Included are explanations and definitions (guidelines for choosing appropriate file formats are provided), sample sites, and a list of references (http://www.columbia.edu/acis/dl/imagespec.html). The Berkeley SUNsite (http://sunsite.berkeley.edu/) has long been regarded as both an excellent source of resources, technical and otherwise for digital endeavors as well as an example of a successful digital project.

Conference proceedings

Perhaps the most insightful material to emerge in the last several years has been presentations from annual conferences such as *Internet Librarian* or *Computers in Libraries* (sponsored by Information Today). These works range in depth and content. They include lessons learned, findings from industry experts, miniature case studies, project overviews, informational presentations, new or early thinking, and practical tips.

Conference literature tends to follow current themes. Most recently, both development of and content management for corporate intranets are dominant. Though the proceedings were not yet available at the time this paper was written, presentation descriptions for the *Internet Librarian '99* conference (San Diego) betray this trend: "The Systems Librarian and Corporate Intranet IT"; "Intranets that Work - Why Ours Didn't, How we Fixed Them, and How You can Fix Yours"; and "Killer App: Library & Finance Intranet Product."

Past conferences offer several interesting case studies of digitization or related efforts to make content available electronically. Other relevant presentations come from professionals, consultants, or firms that have special expertise in building, managing or supporting digital collections.

One such firm insists that a digital library must always be grounded in users needs and function, must use mature technology, and must minimize the burden on users (Wheeler and Rother, 1999). Further, they maintain that the techniques that librarians have developed and time-tested are key. Their list of guidelines and issues reflect this approach and include developing clear acquisition, collection development and management policies, cataloguing techniques, and classification schema.

One large corporation that has struggled with issues in digital content is Boeing, with more than 200,000 employees in the United States alone (Crandall 1998). In his presentation at the 1998 *Internet Librarian* conference, Mike Crandall provides a detailed account of how the organization has approached and developed electronic content. Key to Boeing's success have been user studies that helped shape the overall infrastructure and design. In his estimation, there is a need to better standardize field descriptors (also known as meta tags) as well as search and other protocols that would enable sharing information.

The lack of standardization is noteworthy, since it is often referenced in the conference literature. Most recently, it was identified at the fourth Digital Library conference (Berkeley, California, August 1999) sponsored by the

Association for Computing Machinery. Presenters there noted the disjointed and often competitive efforts to establish standards for metadata and communication (Feldman 1999). Those embarking on a digitization project will find it both a source of frustration and an opportunity to influence.

How-to Manuals and Process Guides

Recently, process guides and manuals have emerged that are targeted for digital librarians, "cybrarians," webmasters, and the like. Formats generally are print, but a few are available online or have accompanying CD-ROMs. This body of work represents the "nuts and bolts" of digitizing. However, because technology changes rapidly, some of these guides cannot be expected to have a long shelf life.

In this genre of work, two major reports of painstaking detail stand out: William Saffady's "Digital Concepts and Technologies for the Management of Library Collections: An Analysis of Methods and Costs" (1995), and David Barber's "Building a Digital Library: Concepts and Issues (1996), both published in *Library Technology Reports*. Though designed primarily for public libraries, they remain the only soup-to-nuts comprehensive guides for digital library projects. In an era of rapidly obsolescing technologies, the content is enduringly relevant. Together they offer outstanding bibliographies, worksheets for estimating conversion costs , definitions, guidelines and process descriptions. These should be considered "must reads" for anyone, in any type of organization, that is undertaking a digital conversion project.

A number of resources take the form of actual, step-by-step manuals. Some are devoted to the specific tasks or roles that digital library might require. The role of the webmaster, for example, is of growing importance and may include responsibility for the interface that makes digitally published content available (van der Walt 1997). A manual for the Webmaster (Champelli and Rosenbaum 1997) provides both a teaching aid (including a CD-ROM) and addresses intellectual property issues. One of the most recent details the steps one library took to become "virtual" (Stielow 1999).

Committee Reports and Fact-Finding Missions

In anticipation of digitization projects, some government and academic institutions have invested a substantial amount of effort in investigating the various technological and organizational considerations or requirements. A number of these have produced reports with recommendations that are useful for others embarking on a digitization project.

Observations made by participants on committees at the onset of large-scale digitization projects offer interesting insights still relevant, though they were made some time ago. For example, corporate executives from technology-based corporations who were asked to participate in workshops in anticipation of building the NSF digital library were asked what they believed the "grand challenges" of the future digital library would be (NSF Source Book 1993). Indexing, navigation, and making materials reachable were consistent themes. Participants then, as now, recognized the need to carefully define "document"

before digitizing its content. For example, does document mean only content, or does it include structure and images? The need for platform independent solutions that enable the sharing of documents across all systems (e.g., PDF) and technologies was also recognized

Another rich source of information valuable to digitizers includes a useful discussion of the lifecycle of digital information (Preserving Digital Information 1996). Information objects in digital form are "created, edited, described and indexed, disseminated, acquired, used, annotated revised, re-created, modified and retained for future use or destroyed . . ." The report includes cost estimations for digital vs. depository libraries that, with some modification, can be applied within a business setting.

When establishing the objectives for digitizing (or as they preferred to call it, imaging) at Penn State University, the task force formed at the outset recognized that despite its advantages, imaging was not a solution to all problems. They recognized that major change would be required, both in culture and environment (Imaging for Process Improvement 1995). Other key aspects of an imaging project the task force identified were the need to acknowledge information as a shared asset, the importance of timely access, reduced dependency on paper files, the need for placing high value on training and education of employees, and a respect for privacy and copyright. Most of these issues are arguably as relevant to the private sector.

Perhaps the most practical and easiest to use reports, however, was prepared for the Washington State Library Council early this year. It identifies the issues, components, computer requirements, and operational issues among other items that resulted from digitizing the Washington State Library. While not comprehensive, it offers an excellent resource for understanding what is involved in a digitization effort.

(http://www.statelib.wa.gov/projects/Digitize/Digitization10.html)

CHAPTER III

METHODOLOGY

This investigation uses no one methodology exclusively. It is in part a real case study of the author's personal experience in digitizing a content-specific collection for a firm that is highly knowledge-dependent. Some informal interviews with others in the organization undergoing similar text-to-electronic conversions were also conducted. However, because many of the typical decisions needed in a digitization project (e.g., hardware, software, document delivery system) were already predetermined by existing internal standards, significant research was spent sifting, sorting, and analyzing literature to supplement this experience.

In addition to the literature search, attempts were made to gather personal experiences of corporate information professionals who had gone through a digital conversion process. Response was poor. Only one resulted in finding meaningful material, and it became apparent that conducting a case study outside my own organization would not be possible.

There are several possible explanations for this. One, I suspect that those who take time to document their experiences for others are inclined to share them via conferences, workshops, or other forums rather than freely share them for inclusion in someone else's research. Second, I believe that many corporations, my own included, do not want to expose their infrastructure for fear of revealing competitive advantages.

Despite these obstacles, it was possible to synthesize results into frameworks that effectively represent best practices. Criteria used to determine what was "best" include the following: frequency of being mentioned; applicability to various settings; independence from underlying assumptions; validated by literature; and based on experience.

In addition to these frameworks, the need for useful tools emerged. While not comprehensive, an attempt was made to develop resources, such as an evaluation guide for scanners, that might be helpful. Obviously, a project of this nature incurred limitations on the breadth and scope of these materials.

CHAPTER IV

FINDINGS AND END PRODUCTS

Key findings and end products are detailed below. Summarized, these include:

1. Best practices, or the "eight best habits of highly effective digitizers." (Appendix 1). This list represents a synthesis of corporate case studies, other relevant literature and the author's own personal experience.
2. Tools, developed from external sources or original, that can serve as aids in a digitization project. Included are digitization project checklist (adopted from the Library of Congress), a sample scanning process flow, and typical elements of an image processing (digitization) system (Appendixs 2, 3, and 4).
3. Scanner and OCR (Optical Character Recognition) guides. (Appendices 5 and 6)

Best practices

What are some of the characteristics of the best digitization efforts? What will ensure the digitization project is successful? Many factors are at play. They can be distilled into eight general categories, below (summarized in Appendix 1).

1. Be on the cutting edge, not the bleeding edge of technology. No organization has bottomless resources. New technologies and products are constantly emerging with promises of delivering information better, faster, easier. The best digitization efforts, however, use technologies and products that

have a proven market base. Even well-established vendors must be approached with caution. Several years ago in an effort to reinvent self, Xerox "DocuTechs" system promised to transform the publishing industry with its copy/digitize/print/transmit capabilities (The Economist, 1994). Had corporations made the $250,000+ investment, they would be sadly disappointed today. This does not suggest that innovation is not desirable. It simply cautions that risks to the overall organization need to be carefully evaluated should the technology or product not live up to expectations.

2. Knowledge management is not left out. Knowledge management gurus, such as Peter Drucker, have been saying this for years. The best technology will not mean a thing unless the organization learns about *how* it learns. Data and document management tools can significantly augment knowledge management tools (Ruggles 1997). Knowledge management software is still in its infancy, however, and users needs to know that terminology will vary from vendor to vendor (Liebowitz 1999). Not every digitization effort requires a highly developed and articulated management strategy. However, even the smallest project should serve as an opportunity for the organization to examine its overall knowledge development strategies.

3. People are key, not the technologies. Over and over, corporate efforts to digitize or create intranets were ultimately reliant on the well-honed skills of – you guessed it – information specialists (NSF Source Book 1993; Henning 1998; Crandall 1998). Why? Information professionals bring

organization to chaos, understand issues of access, and develop the taxonomies and structure that make finding information easier.

It is worth mentioning that the literature supports the case that the role of the information specialists is even more, not less, important than ever. In a recent speech, SLA President Judy Field stated that today's complex information environment is driving a need for creative, innovative specialists (Field, 1998). Not more than five years ago - when Web technology was at a near frenzy - at least one large corporation (Ford Motor) recognized the importance of the specialist as intermediary, a role that still continues to be important. (Schwarzalder 1995).

4. End users needs must be considered. The death of any digital collection is that it does not get used. In many documented cases, the fault is not that the material was not useful, but that it was difficult or even impossible to find. When it comes to finding information, "keep it simple or don't bother" seems to best sum up successful user interfaces, especially for knowledge-dependent organizations (Foy, 1998).

Again, information specialists can play a key role in developing simple, directory-like structures that are intuitive to navigate. Usability studies have also proven effective in isolating and correcting major obstacles to usage. One such study found that most end users within the corporation found information more effectively by browsing than searching (Hennig 1998). Any digitization

project needs to take into consideration how the ultimate recipients of the information are going to access it.

5. Budget for communication and training. Communication is essential throughout the entire lifecycle of a digitization effort. Training is equally important, and often the most neglected component (Crawford 1998). Communication may be strategic (i.e., to influence key decision makers) or informational, but both time and money need to be budgeted for it in any digitization project. Communication, as one corporate executive said a long time ago, should always be the ultimate objective (Blue 1984). One academic library manager named it one of the 5 "Cs" of managing the digital library (Downs 1999). Careful and planned communication helps you obtain necessary buy-in, deflect resistance, and ensure accountability. Communicate "ad nauseum," says one manager recently oversaw an enormous conversion of paper to PDF for a large corporation, and if you don't know the language of upper management, then learn it (Hulser and Spiegelman 1999).

6. Develop metrics for effectiveness and success. Most corporations operate on cost justification for proposed projects. Digitization is no different. Successful projects anticipated ways to measure and report success, either in terms of improved productivity, access to materials (web tracking), or savings (e.g., space requirements, duplication expenses). A rule of thumb offered by one industry executive fifteen years ago still holds true: You must have an easily identified need to have vast quantities of information close at hand for rapid and

frequent look up. The cost saving of look-up speed should justify the cost of the project (Blue 1984). Metrics can also serve as an ongoing public relations device to proactively manage the expectations of management and the organization. Examples might include reports of increased usage, "best seller" reports for accessed materials, or a regular feature of recent additions.

7. Choose carefully. In a digitization project, there are many choices. A partial list of these is provided below, along with issues to consider. However, the best digitizers chose everything thoughtfully and carefully.

As already established in the literature review, there are excellent sources for guidelines (academic web sites, annotated bibliographies, manuals). These address hardware/software requirements, costs, benefits, file formats, software, and intellectual freedom. Your choices should be based on criteria that fit your organization's needs. There are almost no absolutes here – only the legal ones surrounding intellectual property and copyright restrictions.

DIGITIZATION PROJECT CHOICES	
Area	**Issues to consider**
Whether to digitize	Cost/benefit Access Preservation
What to digitize	Volume Size Format Relevance Age Availability elsewhere Intended customers and use Intellectual property and copyright restrictions

BEST COPY AVAILABLE

DIGITIZATION PROJECT CHOICES (cont'd)	
Area	**Issues to consider**
How to digitize	Hardware (servers, network, scanners, printers) Software (for digitizing, storage, and retrieval) In-house vs outsource File formats and naming conventions

8. Leave clues for your successors. This "best practice" was identified by only one person, but is nonetheless one of the most important. Ten years from now, you may find yourself succeeding an information manager in another organization and be asked to implement a major upgrade. How well you understand what was done and why will be a key factor in your ability to do so. Do not assume that someone ten years from now will know what those thin, shiny discs neatly stacked in a tower are. Document everything you can, and keep things simple so they can be managed simply. In a digitization project, attention to the small details like file-naming conventions and directory structures will ensure the next generation's success. (Sarnowski 1998)

Tools

Digital conversion checklist (Appendix 2): The Library of Congress has developed and refined an extensive checklist that it has used for its digitization project. This checklist is presented in its entirety; however, items that are likely of less relevance in corporate or special libraries are shaded (by the author). While the checklist does not address organizational or infrastructure issues, it is a

useful project management tool and with some modification can easily be converted to a Gantt chart. The importance of project management skills in digitizing efforts is cited as a key success factor by some in digitization projects

Scanning work flow diagram (Appendix 3): This process diagram was developed (by the author) to outline the workflow for scanning a paper-based file collection into PDF format. The process assumes that the preliminary steps of weeding the collection, selecting suitable content, and determining hardware/software requirements have already been made.

In this particular case, the decision to scan was based on volume. The amount of materials greater than 25 pages did not justify outsourcing to a vendor. The scanner selected was again based on current and anticipated need, and the scanning software/OCR (Adobe Acrobat) was the clear choice for producing manageable files that could be delivered by the existing Lotus Notes Domino system.

Elements of an image processing system (Appendix 4): This diagram is a visual representation of the general elements of an image processing system used in digitizing (Lynn-George 1996). In includes some of the key factors that

Scanners and OCR guides (Appendices 5 and 6)

Any digitization project under consideration will likely require a knowledge of scanning and OCR technologies. It is impossible to provide a complete and comprehensive guide within the confines of a research project.

However, these Appendicess provide some basic information as well as offer a perspective on considerations for using these tools.

It must, of course, be remembered that technology is constantly evolving, and scanning technology in particular is difficult to keep up with. The guides include a number of resources that can provide relatively current information.

CHAPTER V

CONCLUSIONS

Results of this investigation point to three major findings. First, there are discernible best practices from a growing body of literature that addresses the issues, concerns, and results of digitization efforts. Though the three basic types of libraries (public, academic, special) have different needs, there is enough commonality to argue that some guiding principles applicable to all.

Second, that being said, it is also apparent that there is no "right way" of building a digital library or digitizing a collection. There are no one-stop-shopping manuals that tell one how to do it, nor are there any vendors who can deliver on the promise of a system that fully integrates web resources with the unique resources that may exist in a variety of formats within an organization. In short, the Utopian digital library does not exist. Technology simply isn't mature enough, nor are information codification and other standards fully in place. The smartest thing an organization can do to build a successful digital collection, it seems, is to recognize that there are many ambiguities.

Finally, there is a notable absence of in-depth research on digitizing and digital projects in the private sector. This is understandable: corporations cautiously guard their strategies and less inclined to share technological success stories that might give them a competitive advantage. Nonetheless, opportunities for further study clearly exist. Some possibilities that the literature seems to suggest include: (1) Measuring the *effectiveness* of a digital library in

meeting stated objectives; (2) Quantifying the degree of influence the Internet has had on determining document delivery standards; (3) Surveying an array of industries to develop a perspective on which ones consider digitizing a major priority and why; (4) Analyzing the stability and longevity of products designed to support digitization efforts.

I add only one other personal observation based in part on this project and on my own experience as an information resources specialist with a spike in technology. In the literature, there is a strong undercurrent – even an assumption that today's information specialists in the private sector better know technology, and know it well. To some degree it implies that the highest value you can provide is to become a network engineer cum PERL script guru cum information specialist.

Poppycock. My personal advice: practice the middle way. Learn what technology does or can do, but not how to do it unless you desire to be a techhead. Your skills as organizer, codifier, and facilitator in the growing "swamp" of information will serve you and – more importantly – information seekers far better than knowing how to tweak that Applet on your library's home page.

APPENDIX 1
EIGHT BEST PRACTICES OF HIGHLY EFFECTIVE DIGITIZERS

1. **On the cutting edge, not the bleeding edge, of technology**
2. **Knowledge management is not left out**
3. **People are key, not the technology**
4. **End users are considered**
5. **Communication and training are a line item in the budget**
6. **Effectiveness is measured, reported, and addressed**
7. **Choices are made carefully and thoughtfully**
8. **Clues are left for future generations**

BEST COPY AVAILABLE

APPENDIX 2

NDLP Project Planning Checklist

Library of Congress, National Digital Library Program

January 1997

This document outlines the production process for historical collections at the Library of Congress and reflects that institution's administrative structure and procedures. Not every collection requires all of the steps listed; some collections require additional steps not listed. In practice, many of the operations are carried out in parallel and not sequentially.

Shaded text represents items that can potentially can be amended or eliminated in a corporate library digitization effort

I. Select a collection for digital conversion
   A. Analyze Collection
      1. Determine scope or extent of digitization (entire or subset?)
      2. Assess status of custodial division processing and housing
      3. Assess the status of access aids (degrees of completion, readiness, & format)
      4. Assess best format, e.g. full text conversion, scanned page images
      5. Assess the physical condition and readiness for scanning
      6. Assess restrictions and copyright
   B. Consensus on collection among custodial div, NDLP team, & Library admin

II. Plan the approach to digitization
   A. Develop method and resource plans for collection preparation & digitization
      1. Develop plan for required processing by custodial division
      2. Develop preservation treatment plan
      3. Complete evaluation of physical condition with recommendations
      4. Determine formats for capture, archiving and presentation
      5. Determine physical size (number of characters, images) & special production requirements
   B. Determine repository requirements
      1. Determine scheme for file name assignment
      2. Register aggregate name for collection
      3. Estimate required storage space for digital collection
      4. Update NDL forecast for storage
      5. Evaluate existing finding aids or bib records and develop plan for access aid
      6. Develop plan for framework
      7. Develop restriction plan & implementation (copyright, terms of gift, publicity and privacy)
         a) Find and record restriction facts at collection level
            (1) Search collection files, copyright records, exchange and gift records...
            (2) Create a narrative "findings" statement

APPENDIX 2, continued

b) Find and record restriction facts at the item level
   (1) Search item files, copyright records, exchange and gift records...
   (2) Create fielded/tagged note in access aid

c) Draft proposal for actions to be taken prior to and at the "release" time
   (1) Pre-release: seek required permissions
   (2) Pre-release: Add notices to all restricted items
   (3) Release-time: provide only local or licensed-site access for restricted items
   (4) Release-time: provide no access to items restricted until a given date

d) Draft restriction statement to accompany online collection

e) Review copyright restrictions
   (1) Review facts
   (2) Draft proposal for actions
   (3) Draft restriction statements with advisors in Copyright Office
   (4) Forward findings, action proposal and draft restriction statements to General Counsel for approval
   (5) Revise action proposal and restriction statements after General Counsel review

f) Implement action plans
   (1) Implement pre-release actions, e. g. , seek & obtain permissions
   (2) Implement release-time actions; provide access to collection

8. Workplan for digitization and access aid completed

III. Produce digital collection and access aid
   A. Process and house collection
   B. Implement preservation treatment plan
   C. Item Capture
      1. Preparation
         a) Prepare targets
         b) Prepare scanning instructions specific to collection
      2. Image Capture
         a) Scan collection
         b) Process scanned images
         c) Review images for quality
         d) Coordinate rework
         e) Notify contractor of acceptance of images
      3. Archive images in repository
      4. Text Capture
         a) Prepare keying instructions specific to collection or batch

APPENDIX 2, continued

b) Mark up and key text
c) Review completed text for quality
d) Coordinate rework
e) Process text into final form

5. Archive text in repository
6. Audio Capture
   a) Create preservation and working copy
   b) Determine sample rate and resolution
   c) Select digital audio file format
   d) Analyze storage requirements
   e) Acquire temporary storage space
   f) Perform analog to digital conversion
   g) Edit digital files removing "dead air" at cue-up points
   h) Perform quality review
      (1) Inspect graphic waveforms for truncation and peak
      (2) Audition percentage of sound files
7. Video capture

D. Access Aid Development
   1. Modify existing finding aid
      a) Develop keying instructions
      b) Photocopy and mark up existing printed aid
      c) Coordinate off-site keying
      d) Mark up according to EAD
      e) Review finding aid for accuracy and completeness
   2. Create new finding aid
      a) Verify final processing and arrangement of collection
      b) Draft finding aid
      c) Mark up according to EAD
      d) Review finding aid for accuracy and completeness
   3. Item-level finding aid (Bib record-style)
      a) Set up Minaret database and import item-level records
      b) Upgrade preMARC or MARC records
      c) Download available preMARC or MARC records
      d) Add supplementary notes into Minaret records
   4. Incorporate basic-level links
   5. Add enhanced-access links or subject terms
   6. Prepare collection-level MARC record for future inclusion in MUMS

E. Access Aid Complete

APPENDIX 2, continued

IV. Store in digital archive
- A. Store files in directories as specified by naming scheme
- B. Register items in URN handle-server (when in use)
- C. Deposit items in digital repository (when in use)
- D. All items stored

V. Create Framework
- A. Draft framework components
- B. Review completed framework components
- C. Create mockup of HTML document
- D. HTML mockup approved
- E. Develop and insert hypertext links
- F. Coordinate search engine link with ITS
- G. Insert final links m
- H. Add graphic enhancements to HTML pages
- I. Mount HTML pages on LCWEB server
- J. Review framework for accuracy and completeness
- K. Framework completed

VI. Assemble digital collection
- A. Store access aids in directories as specified by naming scheme
- B. Register document-style access aids in URN handle-server(when in use)
- C. Deposit document-style access aids in digital repository(when in use)
- D. Generate indexes for related MARC records
- E. Generate indexes for textual items in collection
- F. Prepare customized scripts associated with searching indexes and displaying results
- G. Add relevant viewers to supported configuration for WWW access in reading rooms
- H. Assembly completed

VII. Test and refine
- A. Review assembled collection for accuracy and completeness
- B. Test links
- C. Make any necessary changes
- D. Testing completed

APPENDIX 2, continued

VIII. Release Collection

A. Move HTML pages to production area of LCWEB server

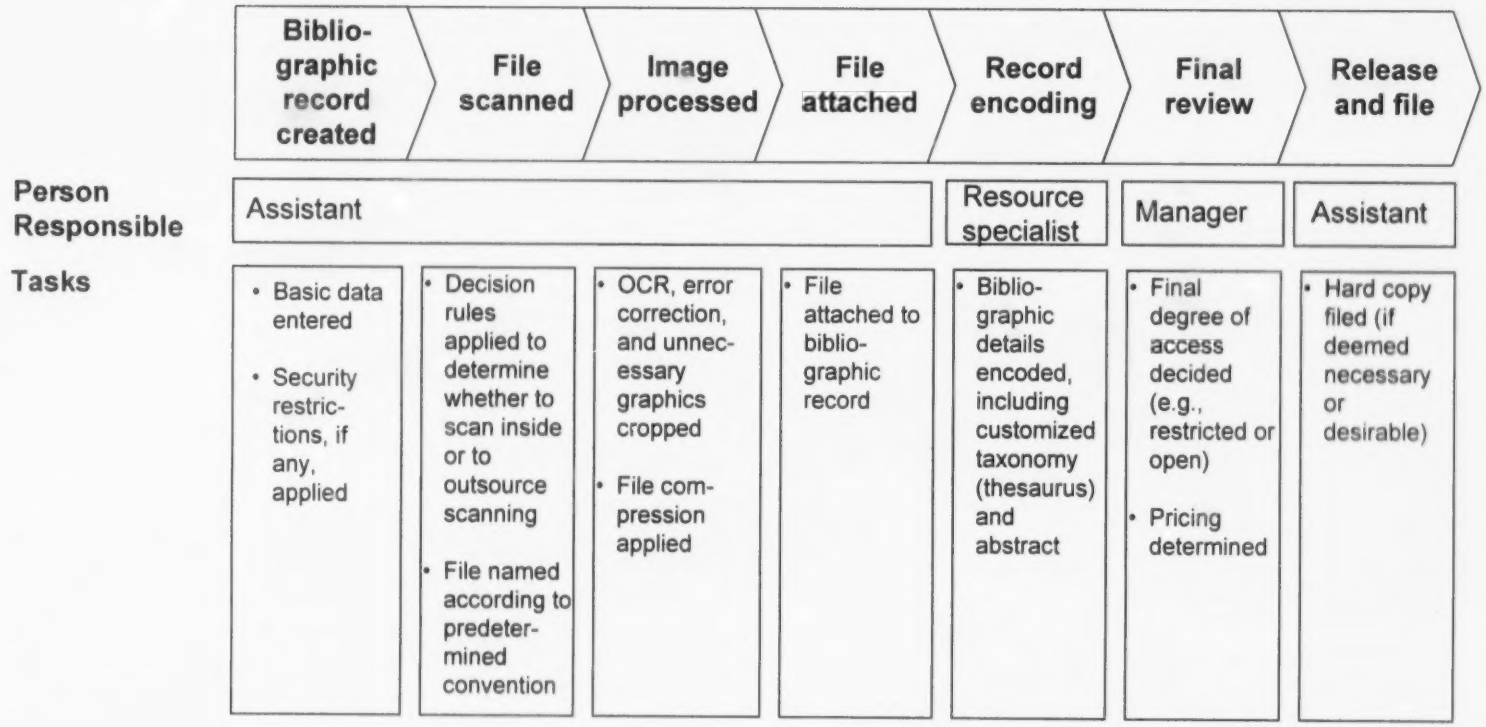
B. Provide links to new collection from appropriate points in LCWEB structure

C. Add collection-level MARC record with pointer in 856 field to MUMS

D. Release digital collection to public

IX. Update

APPENDIX 3
SAMPLE SCANNING PROCESS


Biblio-graphic record created
File scanned
Image processed
File attached
Record encoding
Final review
Release and file
Person Responsible
Assistant
Resource specialist
Manager
Assistant
Tasks
• Basic data entered
• Security restric-tions, if any, applied
• Decision rules applied to determine whether to scan inside or to outsource scanning
• File named according to predeter-mined convention
• OCR, error correction, and unnec-essary graphics cropped
• File com-pression applied
• File attached to biblio-graphic record
• Biblio-graphic details encoded, including customized taxonomy (thesaurus) and abstract
• Final degree of access decided (e.g., restricted or open)
• Pricing determined
• Hard copy filed (if deemed necessary or desirable)

Source: Stach, 1999

APPENDIX 4
ELEMENTS OF A GENERAL IMAGE PROCESSING SYSTEM

	INPUT	PROCESSING	STORAGE	TRANS-MISSION	OUTPUT
Key decisions	Scanning (if yes, then resolution requirements) Manual entry (re-keying) In-house or outsource	OCR Codifying Abstracting Compression	Optical disk Client server Central repository	Acceptable rate and speed of delivery	Monitors Printers
Key influencing factors	Cost Available resources Condition of originals	Intended use User interface Search engine used Storage medium and transmission	File size and volume Technology stability	Input and processing decisions Technical infrastructure (including available bandwidth)	Primary use of digitized objects Quality of digitized material

Source: Lynn-George, 1996

APPENDIX 5: Scanner Guide

(1) HOW DO SCANNERS WORK? THE TECHNOLOGY OF SCANNERS	
Scanners work on the same principle as a camera: light is focused and adjusted through a lens to capture an image. The image is captured on a chemically treated surface and reproduced. The quality of the image is dependent on the equipment used.	SIMILARITIES TO A CAMERA
Instead of a *lens* and *aperture* used to focus and regulate light on a camera, scanners use a *CCD (charged couple device)*. This focuses the image against an array of *electrodes*, which divide the scanned surface into pixels. The image is then focused onto a light-sensitive *silicon substrate*, and the image is then converted to digital format.	LAYERS OF A SCANNER
The degree of "noise," or errors that are captured varies by type of scanner. Higher end scanners reduce this to almost imperceptible levels, while low- to medium-range scanners have higher noise levels.	"NOISE" CONTROL
Scanners do not work by themselves. The *software* which is used (often packaged) with a scanner is instrumental to the quality of the results as well as the ease of use. Most desktop scanners use user-friendly software, while high-end scanners require highly skilled and trained operators.	SOFTWARE

BEST COPY AVAILABLE

APPENDIX 5, continued

(2) WHAT TYPES OF SCANNERS ARE THERE? WHAT ARE THERE ADVANTAGES AND DISADVANTAGES

TYPE	DESCRIPTION		PRICE RANGE (DESKTOP TO COMMERCIAL)
Flatbed	Most common type of scanner used in desktop publishing. Light moves across material, which sits on glass bed, while it is being scanned		$100 to 20,000+
	ADVANTAGES • Cost, availability • Wide range of capabilities • Speed and durability	DISADVANTAGES • Dimensional limitations (of image to be scanned) • No lighting control • Unsuitable for fragile/brittle items	
Sheetfeed	Scanner in which light source remains stationary as material to be scanned is fed across it.		$130-50,000
	ADVANTAGES • High-volume, high-speed	DISADVANTAGES • Dimensional limitations • Limited image enhancement	
Drum	Used most often for graphics arts producers, represent higher end of scanners.		$10,000-30,000
	ADVANTAGES • Highest resolution	DISADVANTAGES • Expensive, slow • Usually 12" x 17" max. image size • Highly expensive to own and operate	

BEST COPY AVAILABLE

BEST COPY AVAILABLE

APPENDIX 5, continued

(2) WHAT TYPES OF SCANNERS ARE THERE? WHAT ARE THERE ADVANTAGES AND DISADVANTAGES		
TYPE	DESCRIPTION	PRICE RANGE (DESKTOP TO COMMERCIAL)
Slide	Used exclusively for phototransparencies (slides, negatives) ADVANTAGES: No size or shape limitation of original image DISADVANTAGES: • Second-generation image degrades quality • Added cost for generating slide • Lesser resolution	$400-17,000
Digital camera	Scanner combined with digital camera; still in early phases of development ADVANTAGES: • Direct scanning from original • No size/shape limitations • Control over lighting • Non-destructive DISADVANTAGES: • Technology is not yet fully market-ready • Slow capture speed • Large file sizes • High operator skill	$28,000-33,000
Microfilm	Used exclusively to convert microfilm or fiche ADVANTAGES: • Good for archival backup • No size/shape limitations of original DISADVANTAGES: • Second or third-generation images • Lesser resolution • Added cost of converting image to microfilm • Expensive equipment	$65,000-240,000
Handheld	Portable, low-end, not useful for quality reproductions ADVANTAGES: • Extremely affordable, easy to use • No size/shape limitations DISADVANTAGES: • Usually limited to black & white or line image capture • No control over lighting	Usually under $300

APPENDIX 5, continued

(3) SPECIAL CONSIDERATIONS FOR INFORMATION PROFESSIONALS	
What will scanner be used for?	*For:* • **Digitizing collection?** • **Graphics design?** • **Text conversion?**
What are the quality requirements and end-user needs?	*Will you need:* • Color, gray-scale, or half-tone? • Low, medium, or high resolution (DPI)? • Readable and/or searchable text • Limited or wide availability of the electronic materials?
What source materials will be scanned?	*Do you have:* • **Fragile or brittle materials?** • **Poor-quality originals?** • **Slides, microfilm, fiche, or other non-standard materials?**
What are the technology requirements?	*Have you considered:* • Size, type of network needed? • Computers you now have, and upgrading them if necessary? • Optimal type of storage media/ devices (e.g., CD-ROM, magnetic disc, "jukebox"?
What skills are needed to perform the task, and how quickly does it need to get done?	*Do you have:* • Qualified in-house expertise, or will someone need extensive training? • Budgetary or time constraints which might make outsourcing a better option?

BEST COPY AVAILABLE

APPENDIX 5, continued

(4) EVALUATING SCANNERS: A FRAMEWORK		
1. SCANNER TYPE *LESS EXPENSIVE* / *MORE EXPENSIVE*		Handheld
2.		Flatbed
3.		Sheetfed
4.		Slide
		Digital camera
		Microfilm
		Drum
5. COLOR SAMPLING/ RESOLUTION		**Black & white sampling (shades of gray) supported**
		Color sampling capability (e.g., 8-bit up to 36 bit, or from 256 colors to billions of colors)
		DPI supported(e.g., 300 up to 2,000)
6. SPEED		**How many pages can be scanned at a time or per minute?**
7. SIZE		**Is the scanner bulky or relatively compact?**
8. NOISE		**Is the scanner noisy or quiet?**
9. BUNDLED SOFTWARE		**What sort of software comes with the scanner? How reliable is it? Is OCR software included?**

APPENDIX 5, continued

10.PAGE SIZES AND TYPES OF MATERIAL HANDLED		**Can the scanner handle assorted paper sizes (e.g., letter or legal)? Can it handle bound materials, loose-leaf, and fragile or brittle materials?**
11.OPERATOR SKILL LEVEL REQUIRED		**What kind of training is required to use and operate the scanner?**
12.OVERALL IMAGE QUALITY		**Are images, whether black and white or color, consistent and of good quality?**
13. PRICE AND WARRANTY		**Is the product well-supported and fairly priced next to comparable models?**

APPENDIX 5, continued

(5) SELECTED RATINGS OF DESKTOP PUBLISHING SCANNERS[1]

CATEGORY: BUDGET SCANNERS

Product	*Price*	*Type*	*Resolution (Real/inter-polated)*	*Software included*	*Pros*	*Cons*	*Verdict*
Visioneer PaperPort Strobe (800) 787-7007	$249	Sheetfed	300 dpi/2,400 dpi	Connectix QuickCards LE, Corex Card Scan, PictureWorks PhotoEnhancer, Visioneer PaperPort, Xerox TextBridge OCR	Tiny, portable unit, excellent software bundle; handles various media sizes, even odd and extra long.	Washed-out color scans; ineffective paper guides mar paper handling; can't scan bound materials	☆☆☆☆ **A sleek and smart contender**
Hewlett-Packard ScanJet 5s (800) 722-6538	$199	Sheetfed	300 dpi/600 dpi	Caere OCR software, Visioneer PaperPort for HP	Good paper handling automatic document feeder	Slow; muddy color and grayscale images; can't scan bound material	☆☆☆ **Showing its age against newer competition**
Info Peripherals ImageReader EPP (800) 777-3208	$130	Flatbed	300 by 600 dpi/4,800 dpi	Caere Recognita OCR, Info Peripherals InfoCenter, Ulead iPhoto Express	Impressive price; extended warranty.	Short cables; lacks document-management software.	☆☆☆ **Outstanding image scanning for the price**
Umax Technologies Astra 600P (800) 777-3208	$199	Flatbed	300 by 600 dpi/4,800 dpi	Adobe PhotoDeluxe, NewSoft Presto PageManager	Fast scans; supports 30bit color and 10-bit gray.	Noisy; software lacks an integrated button bar	☆☆☆ **Get high-end color at budget prices**

BEST COPY AVAILABLE

[1] Selected reviews taken from *PC Computing*, November 1997, Matthew J. Lake and Bob Weibel.

APPENDIX 5, continued

CATEGORY: MIDRANGE SCANNERS

Product	Price	Resolution (Real/interpolated)	Software included	Options	Pros	Cons	Verdict
Agfa StudioStar (800) 227-2780	$749	600 by 1,200 dpi/2,400 dpi	Adobe Photoshop LE, Agfa FotoLook TWAIN driver and FotoTune, Caere OmniPage LE	Transparency adapter, $250; document feeder, $275	Fastest; best image quality; most versatile software; good high-res-zoomed preview	Setup isn't straightforward	☆☆☆☆☆ **Your top choice for quality and image**
Umax Technologies Astra 1200S (800) 562-0311	$649	600 by 1,200 dpi/9,600 dpi	Adobe Photoshop, NewSoft Presto PageManager with OCR	Transparency adapter, $199	Excellent color quality; easy installation.	No high-res zoomed preview	☆☆☆☆ **Almost as good as the StudioStar for $100 less.**
Microtek Lag ScanMaker E6 Pro (800 654-4160	$559	600 by 1,200 dpi/4,800 dpi	Adobe Photoshop, Caere OmniPage LE, Xerox TextBridge Pro	Transparency adapter, $299; document feeder, $400	Good image quality; high-res zoomed preview; nice price.	Slow on 600-dpi scans; tricky cropping	☆☆☆☆ **Strong all the way around except for speed.**
Plustek OpticPro 9630P (800 685-8088	$199	600 by 1,200 dpi/9,600 dpi	ExperVision TypeReader, Micrografx Photo Magic	None	Amazing price; easy software and hardware installation.	Slow scanning; unreliable operation; tricky cropping	☆☆ **Low price, but slow and quirky**

BEST COPY AVAILABLE

APPENDIX 5, continued

CATEGORY: HIGH-END SCANNERS						
Product	*Price*	*Resolution (Real/inter-polated)*	*Software included*	*Pros*	*Cons*	*Verdict*
Epson America Expression 636 Pro (800) 463-7766	$1,399	600 by 1,200 dpi/4,800 dpi	Adobe Photoshop LE, Caere OmniPage LE, Claris HomePage, Epson SilverFast, Meta-Creations Kai's Power Tools, NewSoft Presto PageManager LE, Xerox TextBridge Pro 96	Best image quality; fastest preview and 300-dpi scans; great software bundle; great price	Slow 600-dpi scanning	☆☆☆☆☆ **Unbeatable price and 300-dpi speed, plus software that'll make anyone a graphics pro**
Microtek lab ScanMaker III (800) 654-4160	$1,499	600 by 1,200 dpi/4,800 dpi	Adobe Photoshop, Microtek ScanWizard TWAIN driver Xerox TextBridge Pro	Fast; high color quality; nice price; best for scanning slides.	Tricky hardware setup; inaccurate cropping in zoomed preview	☆☆☆☆ **Solid all around, except software needs polish**
Umax Technologies PowerLook 2000 (800) 562-0311	$3,595	1,000 by 2,000 dpi/10,000 dpi	Adobe Photoshop, Binuscan PhotoPerfect Master, Umax MagicMatch and MagicScan	High resolution for image enlargement; good color quality; easy batch scanning	Inaccurate cropping in zoomed preview with no high-res preview; pricey	☆☆☆☆ **You'll pay for its super high resolutions**

APPENDIX 5, continued

SCANNERS – A SELECTED BIBLIOGRAPHY

Books

Beale, Stephen & James Cavuoto. *The Scanner Book: A Complete Guide to the Use and Applications of Desktop Scanners.* Torrance, California: Micro Publishing Press, 1989.

Day, Jerry B. *Color Scanning Handbook: Your Guide to Hewlett-Packard ScanJet Color Scanners.* Upper Saddle River, New Jersey: Prentice Hall PTR, 1997.

Green, William B. *Introduction to Electronic Document Management Systems,* Boston: Academic Press, Inc., Harcourt Brace Jovanovich, 1993.

Harmon, Craig K. *Lines of Communication: Bar Code and Data Collection Technology for the 90s,* Peterborough, New Hampshire: Helmers Publishing, Inc., 1994.

Hornstein, Jonathan. *Scanning: Your Personal Consultant.* Emeryville, California: Ziff-Davis Press, 1995.

Kenney, Anne R., and Stephen Chapman. *Digital Imaging for Libraries and Archives.* Ithaca, New York: Department of Preservation and Conservation, Cornell University Library, 1996.

Periodicals

Goddard, John & Weibel, Bob. "Scanners Across the Spectrum," *PC World* 13(September 1995): 148-149+.

Lake, Matthew J. & Weibel, Bob. "Scanner Superguide." *PC World* 10 (November 1997): 312-313+.

WWW

http://www.zdnet.com/products/filter/guide/0,7267,1500124,00.html
ZDNet (computer shopping site). Includes fairly inclusive reviews of scanners, prices, capabilities.
http://banking.com/article4.asp
"Bill's Corner." Useful articles discussing evolution of imaging technology. Includes links to developing an imaging plan and where to shop.

APPENDIX 6: OCR (Optical Character Recognition) guide

HOW OPTICAL CHARACTER RECOGNITION (OCR) WORKS	
1. Original is transferred to electronic image, usually by scanning.	IMAGE CONVERTED
2. **The OCR software first acquires the image and establishes its correct orientation based on the text. If necessary, the image is "rotated" from a vertical to horizontal orientation (or vice-versa) before recognizing text.**	**IMAGE ACQUIRED, ORIENTED**
3. Most OCR programs will then attempt to discern the layout of the document, identifying graphics or tables. Text will then be "zoned" into regions for OCR.	ZONES DEFINED
4. OCR is applied. Both characters and words are recognized using OCR and *ICR*, Intelligent Character Recognition.	**OCR AND ICR APPLIED**

BEST COPY AVAILABLE

APPENDIX 6, continued

OCR CHALLENGES	
Problem	Example
Contrast	Text which appears *too light* or **too dark** to be read correctly
Mixed languages	A document containing mixed languages or a mixture of Latin and non-Latin characters (e.g., "Ατηενα, the Greek protectress of Athens")
"Noise"	Spots, stains, or other discolorations which may be incorrectly interpreted
Non-standard fonts or font size	Font sizes which are extremely small, or ornamental or specialty fonts with elaborate serifs

BEST COPY AVAILABLE

APPENDIX 6, continued

OCR SOFTWARE: A SIMPLE COMPARISON OF SELECT PACKAGES

Software	*Price**	*Pros/cons*	*Rating*
Presto	$	**Supports use on network. Can transfer simple, multi-column layouts to Word, Excel, and WordPerfect formats. Poor handling of tables and poor documentation.**	☆☆☆
OmniPage	$$$	**Handles multiple languages. Can transfer to various word processing formats. Excellent documentation. Very expensive for only slightly better performance than competitors.**	☆☆
TextBridge	$	**Comparable to OmniPage, but much less expensive. Good documentation and good value**	☆☆
Acrobat	$$	**Supports many foreign languages and font faces; handles non-standard layouts well. Built-in, line-by-line editing of word "suspects" (unrecognized text). Documentation is not terrific, and converted format is proprietary, cannot be transferred to other WP programs.**	☆☆☆☆

*KEY
$=Under $100
$$=At least $100, but less than $300
$$$=More than $300

☆ ☆ Fair
☆ ☆ ☆ Good
☆ ☆ ☆ ☆ Excellent

APPENDIX 6, continued

OCR and TEXT CONVERSION: SPECIAL CONSIDERATIONS FOR INFORMATION PROFESSIONALS	
What are the special content-related issues?	• **Are there any foreign languages which need to be supported?** • **Is it important to preserve the layout and graphics of the original?** • **Are there unique fonts or special characters which could be problematic?** • **Is copyright infringement an issue?**
What are the end-user needs?	• What degree of manual intervention and correction is acceptable? • Will use of the materials be restricted locally, or will they be made available beyond the physical boundaries of the institution? • Is it important that the material be searchable? • Will you allow the items to be downloaded or printed?
What source materials will be scanned?	• **Are there any fragile or brittle materials?** • **Is the quality of the originals suitable for scanning (e.g., are they stained or excessively light or dark)?** • **Are there slides, microfilm, fiche, or other non-standard materials?**

BEST COPY AVAILABLE

APPENDIX 6, continued

OCR and TEXT CONVERSION: SPECIAL CONSIDERATIONS FOR INFORMATION PROFESSIONALS	
What are the technology requirements?	• What size and type of network do you have? • What kind of computers do you now have, and is an upgrade necessary? • How do you plan to store the digitized materials (e.g., on CD-ROM, magnetic disc, "jukebox")?
What skills are needed, and what are the time constraints?	• **Will the project require extensive training, either immediate or ongoing, or both?** • **Are there budgetary or time constraints which might make outsourcing the conversion a better option?**

APPENDIX 6, continued

PDF (Portable Document Format – Adobe Acrobat software): ADVANTAGES AND DISADVANTAGES	
ADVANTAGES	***DISADVANTAGES***
• Platform and device independent • Special compression technology significantly reduces file sizes, especially graphic (GIF or JPEG) images. • Offers superior design control: enables publishing web sites with complex layouts, high graphic capabilities, hyperlinking, bookmarks, and forms. • Full-text indexable, searchable	• ***Requires Adobe Reader to view*** • ***Cannot be "backwards" converted to original format*** • ***Scanned files can still be too large for the Web*** • ***Software not always the most intuitive to use***

APPENDIX 6, continued

OCR AND PDF: A SELECT BIBLIOGRAPHY

Books

Beale, Stephen & Cavuoto, James. *The scanner book: a complete guide to the use and applications of desktop scanners.* Torrance, California: Micro Publishing Press, 1989.

Harmon, Craig K. Harmon. *Lines of communication: bar code and data collection technology for the 90s,* Peterborough, New Hampshire: Helmers Publishing, Inc., 1994.

Hornstein, Jonathan. *Scanning: your personal consultant.* Emeryville, California: Ziff-Davis Press, 1995.

Image processing and optical character recognition – how they work and how to implement them, New York, New York: American Institute of Certified Public Accountants (AICPA), 1993.

Kenney, Anne R., and Stephen Chapman. *Digital Imaging for Libraries and Archives.* Ithaca, New York: Department of Preservation and Conservation, Cornell University Library, 1996.

Murshed, Nabeel A. and Flávio Bortolozzi, eds. *Lecture notes in computer science 1339: advances in document image analysis,* New York: Springer, 1997.

Ogg, Harold C. and Marlene H. Ogg. *Optical character recognition: a librarian's guide,* Westport, Connecticut: Meckler, 1992.

Schantz, Herbert F., comp. *Recognition technology in the information industry – 1992 NFAIS Report Series*, Philadelphia, Pennsylvania: The National Federation of Abstracting and Information Services, 1992.

Siegel, David. *Creating killer Web sites: the art of third-generation site design*, Indianapolis, Indiana: Hayden Books, 1996.

Warner, J.; Milburn, K.; & Burdman, J. *Converting content for Web publishing: time-saving tools and techniques,* Indianapolis, Indiana: New Riders Publishing, 1996.

Waters, Crystal. *Web concept & design – a comprehensive guide for creating effective Web sites*, Indianapolis, Indiana: New Riders, 1996.

APPENDIX 6, continued

Periodicals

"GATF Encyclopedia, PDF Bible," *Editor & Publisher,* XX(March 7, 1998), 27.

Lorenz, Lonn. "PDF promises speed," *Folio*, XX(??), 54.

"Presto 3.0: Quality OCR at a bargain rate," *PC World*, XX(January 1998), 128.

WWW

http://www.onix.com/tonymck/INDEX.HTM
"Online OCR Lab" offered by consultant. Includes discussion of OCR, text searching, and digital libraries.

http://www.adobe.com/prodindex/acrobat
Adobe Corporation. Product description of Acrobat, used to create PDF files.

http://www.tumbleweed.com/solutions/ime_overview_archwhitepaper.htm
Tumbleweed Communications (commercial vendor of document delivery systems). White paper on architecture of document delivery. Intended to sell the benefits of this vendor's product, but informative and useful for understanding architecture.

REFERENCE LIST

Aramburu Cabo, Mara Jose and Rafael Berlanga Llavori. 1997. An approach to a digital library of newspapers. *Information Processing & Management* 33 (September): 645-61.

Association of Research Libraries. 1995. Definition and purposes of a digital library. Unpublished (October 23). http://www.ifla.org/documents/libraries/net/arl-dlib.txt

Ballard, Terry. 1996. Library systems: Keeping up our images. *Information Today* 13 (April): 52-53.

Barber, David. 1996 Building a digital library: concepts and issues. *Library Technology Reports* 32 (September/October): 573-738.

Barber, David. 1998. Tools for managing the digital library: guidelines and sample RFPs. *Library Technology Reports* 34 (July/August): 439-551.

Barbera, Jose. 1996. The intranet: A new concept for corporate information handling. *Online Information 96. Proceedings of the International Online Information Meeting.* (London, England, United Kingdom) December 3-5.

Beagrie, Neil and Daniel Greenstein, eds. 1998. A strategic policy framework for creating and preserving digital collections. Arts and Humanities Data Service (July 14) http://ahds.ac.uk/manage/framework.htm

----------. 1996. Bibliography on electronic records: organizations and business processes. Unpublished . http://www.lis.pitt.edu/~nhprc/bib3.html

Blue, Roger E. 1984. President's Message. *Journal of Information and Image Management* 17 (September): 22-23.

Bowes, Roger. 1995. How best to find and fulfil business information needs. *ASLIB Proceedings* 47 (May): 119-26.

Connor, Louis. 1995. Divide, conquer and digitize. *Communications Week* 574 (Sep. 11): S7-S14.

Crandall, Mike. 1998. Issues in digital content delivery to end users. . Internet Librarian '98 - Proceedings—1998. Medford, New Jersey: Information Today, Inc.

Crawford, Gregory A. 1999. Issues for the digital library. *Computers in Libraries* 19 (May): 62-4.

Crawford, Walt. 1998. The danger of the digital library. *The Electronic Library* 16 (February): 28-30.

Downs, Robert. 1999. The 5 C's of managing the digital library. 1999. 14th Annual Computers in Libraries - Proceedings—1999. Medford, New Jersey: Information Today, Inc.

Ensor, Pat. 1997. The cybrarian's manual. Chicago and London: American Library Association.

Field, Judy. 1998. Information + technology + you equals knowledge management. 13th Annual Computers in Libraries - Proceedings—1998. Medford, New Jersey: Information Today, Inc.

Forrest, Moyra. 1998. Towards the digitial library. *Library Association Record* 100 (October): 540.

Fox, Edward, ed. 1993. Source book on digital libraries. National Science Foundation. December 6, 1993.

Griffith, Richard. 1997. Yes, they scan. *Office Systems* 14 (April): 41-45.

Helal, Ahmed H. and Joachim W. Weiss, eds. 1996. Towards a worldwide library: a ten year forecast. Essen [Germany]: Universitätsbibliothek Essen.

----------. 1995. Imaging for Process Improvement: Report of the Imaging Committee [at Penn State University]. Unpublished (July) http://www.psu.edu/computing/imaging.html

Hennig, Nicole. 1998. Going forward: usability testing the web site. . Internet Librarian '98 - Proceedings—1998. Medford, New Jersey: Information Today, Inc.

Hulser, Richard and Barbara Spiegelman. 1999. Essential technologies for going digital: a workshop. (Unpublished): March

Kilker, Julian and Geri Gay. 1998 The social construction of a digital library: a case study examining implications for evaluation. *Information Technology and Libraries* 17 (June): 60-70.

Kuny, Terry and Gary Cleveland. 1998. The digital library: myths and challenges. *IFLA Journal* 24 (March): 107-13.

----------. 1999. Issues in digitization: a report prepared for the Washington State Library Council. http://www.statelib.wa.gov/projects/Digitize/Digitization10.html

Lee, Sul H. 1997. Economics of digital information collection, storage and delivery. New York: Haworth Press.

Lehtonen, Eeva-Liisa and Teppo Savinen. 1997. How the virtual business library meets the virtual researcher - a case of a customer driven innovation. *INSPEL* 31 (1997): 204-12

Lesk, Michael. 1997. Going digital. *Scientific American* (March). http://www.sciam.com/0397issue/0397lesk.html

Liberman, Kristen and Jane L. Rich. 1993. Lotus Notes databases: The foundation of a virtual library. *Database* 16 (June): 33-40.

Liebowitz, Jay, editor. 1999. Knowledge management handbook. Boca Raton: CRC Press.

Lynn-George, Jann. 1996. Digitization: a literature review and summary of technical processes, applications and issues. May 10, 1996. http://www.library.ualberta.ca/library_html/libraries/law/digit1.html

Lynn-George, Jann. 1996. Digitization: technical processes, applications and issues, a selected annotated bibliography. May 10, 1996. http://www.library.ualberta.ca/library_html/libraries/law/digit2.html

Manville, Brook. 1993. Tradition and innovation in the management of professional knowledge: a case study of a "virtual library." *Harvard Library Bulletin* (Spring).

Manville, Brook and Nathaniel Foote. 1996. Strategy as if knowledge mattered. *Fast Company* 2 (April): 66. (http://www.fastcompany.com/online/02/stratsec.html)

Megill, Kenneth A. 1997. The corporate memory: information management in the electronic age. London: Bowker Saur

Nath, Ravinder. 1994. Difficulties in matching emerging information technologies with business needs: A management perspective. *Information Processing & Management* 30 (May/June): 437-44.

Nixon, Carol and Heide Dengler. 1999. 14th Annual Computers in Libraries - Proceedings—1999. Medford, New Jersey: Information Today, Inc.

Nixon, Carol and Heide Dengler. 1998. 13th Annual Computers in Libraries - Proceedings—1998. Medford, New Jersey: Information Today, Inc.

Nixon, Carol, M. Heide Dengler and Mare L. McHenry. 1998. Internet Librarian '98 - Proceedings—1998. Medford, New Jersey: Information Today, Inc.

Powell, Alan. 1994. Management models and measurement in the virtual library. *Special Libraries* 85 (Fall): 260-3.

----------. 1996. Preserving digital information: report of the task force on archiving digital information. Commissioned by The Commission on Preservation and Access and The Research Libraries Group. May 1, 1996.

Ruggles, Rudy. 1997. Knowledge tools: Using technology to manage better (working paper). Ernst & Young (April) http://www.businessinnovation.ey.com/mko/html/toolsrr.html

Ryan, William. 1989. My First Wish Is for a Fast New Scanner. *Computer Technology Review* 9 (March): 14, 21.

Saffady, William. 1995. Digital library concepts and technologies for the management of library collections: an analysis of methods and costs. *Library Technology Reports* 31 (May/June): 221-380.

Sarnowski, John. 1998. Why digital will fail and what we can do about it. Internet Librarian '98 - Proceedings—1998. Medford, New Jersey: Information Today, Inc.

Schwarzwalder, Robert. 1995. The Sci/Tech Image Invasion: Approaches to Managing the Digital Library. *Database* 18 (August/September): 81-84.

Scott, R. Neil. 1991. The importance of business and competitor intelligence (BCI) for effective corporate strategy and planning. *The Southeastern Librarian* 41 (Spring): 12-14.

Stielow, Frederick, editor. 1999. Creating a virtual library. New York: Neal-Schuman Publishers Inc.

St. Clair, Guy. 1993. The future challenge: management and measurement. *Special Libraries* 84 (Summer): 151-4.

Tennant, Roy. 1998. Digital library infrastructure. *Library Journal* 123 (June 15): 32+.

van der Walt, Pieter W. and Pieter A. van Brakel. 1997. The webmaster: a new player in the information centre's online team. *Electronic Library* 15 (December): 447-54.

Van Gils, Gilbert. 1994. How a government company improves its efficiency by adopting a digital storage and retrieval system. *IMC Journal* 30 (May/Jun): 10-11.

Wheeler, Barry and Barbara Rother. 1999. Creating personal digital libraries. 14th Annual Computers in Libraries - Proceedings—1999. Medford, New Jersey: Information Today, Inc.

Wilson, Thomas D. 1997. Information behaviour: An interdisciplinary perspective. *Information Processing & Management* 33 (July): 551-72.

Woods, John. 1984. OCR has a place in data processing. *Data Processing* 26 (June): 10-12.

----------. Xerox: Well documented. 1994. *The Economist* 333 (Oct 1): 86-89.